Diplomatura en Análisis de Datos aplicados al Desarrollo de Políticas Públicas

Título del tema de trabajo Final:

Programa Cambio Rural – SAGyP

Integrantes

Constanza Guerrini
Patricia Perrone
Marianela Pi

Tabla de contenido

Objetivos Generales

El objetivo general de esta investigación es caracterizar a los beneficiarios/as del Programa Cambio Rural.

Cambio Rural es una política pública que busca, a través de la asistencia técnica, promover y facilitar la intensificación y reconversión productiva, como un medio para mejorar la situación productiva y socioeconómica de los pequeños y medianos productores rurales y propender al desarrollo agroindustrial en todo el territorio nacional, impulsando el aprendizaje grupal.

Objetivos Específicos

Obtener un *dataset* limpio, determinar las variables a utilizar para la caracterización y realizar la clusterización. Realizar análisis estadísticos.

Antecedentes:

La situación actual es la reestructuración del Programa Federal de Reconversión Productiva que tuvo lugar en septiembre de 2017 y que fue relanzado como Programa Cambio Rural (CR) a través de la Resolución E 249/2017, junto con un nuevo Manual Operativo. En ese mismo documento se crea el Registro de Integrantes de Grupos Cambio Rural y se desarrolla un Sistema de Gestión para sistematizar la información de este Registro. La decisión de reestructuración se tomó debido a que, a pesar de los esfuerzos y recursos invertidos, los resultados obtenidos no fueron suficientes para posicionar a la pequeña y mediana empresa rural en los niveles óptimos y necesarios de eficiencia productiva que les permitieran enfrentar exitosamente las fluctuaciones económicas y climáticas.

Como antecedente, se tomó la Resolución N° 227 de fecha 4 de mayo de 1993 de la Secretaría de Agricultura, Ganadería y Pesca del entonces Ministerio de Economía y Obras y Servicios Públicos, que creó el Programa Federal de Reconversión Productiva con el propósito de promover y facilitar la intensificación y reconversión productiva de la pequeña y mediana empresa rural. La creación del programa se llevó a cabo en el contexto histórico de los años 90 y ante la crisis económica reinante, y se solicitó propuestas al sector, una de las cuales, elaborada por el INTA, se convirtió en el Programa Federal de Reconversión Productiva. En su origen, el Programa otorgó gran importancia al trabajo coordinado con las Provincias y las entidades del sector, incluidas las intermedias, para posibilitar y potenciar la asistencia técnica, el acceso al crédito y el intercambio tecnológico necesario para una mayor eficiencia y diversificación productiva que, junto al esfuerzo asociativo, generaran economías competitivas.

Políticas similares

https://agriculture.ec.europa.eu/common-agricultural-policy/rural-development_es

Actividades y metodología:

a) Ver las variables relevadas
b) Selección de las variables a trabajar
c) Limpieza de datos
d) Homogeneización de datos
e) Generación dataset
f) Análisis estadístico de las variables

Factibilidad:

Es posible realizar un análisis de clustering utilizando variables categóricas, pero es importante utilizar una técnica adecuada para manejarlas correctamente. En este caso, utilizaremos la técnica de codificación de variables categóricas, como la codificación "One-HotEncoder", la cual implica la creación de una variable binaria por cada posible valor de la variable categórica original.

Sin embargo, es importante tener en cuenta que trabajar con demasiadas variables categóricas puede generar un gran número de variables binarias en el análisis, lo que puede afectar la calidad del modelo resultante. En este caso, nos enfocaremos en trabajar con un máximo de 4 variables categóricas, ya que abordar más variables requiere un mayor conocimiento y experiencia en la selección de técnicas de clustering adecuadas y en la interpretación de los resultados obtenidos, lo cual puede exceder el plazo de dos meses disponible.

PROCESO

Se realizó un análisis de la tabla y se identificaron varios problemas en las variables de edad y superficie, principalmente por errores de carga.
Se intentó resolver el problema de los datos de la superficie utilizando una transformación logarítmica para reducir el rango de datos, pero no se pudo utilizar por ser una función que no está definida en el cero (dentro de los beneficiarios hay quienes tienen actividades que no requieren de la explotación de tierras por lo que la superficie utilizada tiene un valor genuino de cero). La posibilidad de sustituir por la media no nos pareció adecuada en estas circunstancias. Otro tema del cual ocuparse son los NaN.

Se pensó como alternativa la utilización de una función partida para el uso del logaritmo, donde la función es 0 para x=0; *Null* para x=NaN y Log(x) para x>0. En este caso tendríamos además una “sobrecarga” de ceros, por los ceros genuinos y los provocados por el log(1), generando datos falsos.
Para la edad decidimos utilizar una segmentación del rango etario que transforme los errores de carga en *outlayers (menores de 18, mayores de 99), utilizando sólo las categorías* que van desde mayores a 18 años a menores de 99.
Por todo esto se decidió trabajar con las siguientes variables categóricas, rango etario, educación, ingresos (% de ingresos mensuales que aporta el emprendimiento acompañado por CR) y situación AFIP.

Resumen estadístico de las cuatro variables

	Rango_etario	Educación	Ingresos	SituacionAFIP
count	13391	13391	13391	13391
unique	4	6	3	17
top	Más de 55	Secundario	Más del 50%	Responsable Inscripto
freq	5326	5510	5808	3909

El valor "top" en cada variable muestra la categoría más frecuente en esa variable, lo que indica que hay 5,326 registros con la categoría "Más de 55" en la variable "Rango_etario", 5,510 registros con la categoría "Secundario" en la variable "Educación", 5,808 registros con la categoría "Más del 50%" en la variable "Ingresos" y 3,909 registros con la categoría "Responsable Inscripto" en la variable "SituacionAFIP".

Construcción del Modelo

Nuestro objetivo en este cuaderno es sólo mostrar el algoritmo K-Modes, omitiendo en esta ocasión el EDA y pasaremos directamente a la construcción del modelo.

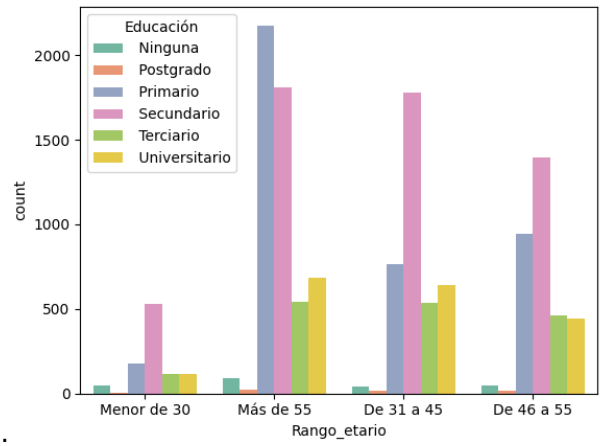
Construcción del modelo - Convertimos las variables categóricas en numéricas y las agregamos a la base original.

	Rango_etario	Educación	Ingresos	SituacionAFIP	Rango_etarioNum	EducacionNum	IngresosNum	SituacionAFIPNum
0	Menor de 30	Ninguna	Menos del 30%	No declarado	0	0	0	0
1	Menor de 30	Ninguna	Más del 50%	Responsable Inscripto	0	0	2	3
2	Menor de 30	Ninguna	Más del 50%	Responsable Inscripto	0	0	2	3
3	Menor de 30	Ninguna	Más del 50%	Monotributo Cat B	0	0	2	6
4	Menor de 30	Ninguna	Más del 50%	Monotributo Cat C	0	0	2	7
...	...	...	...	...	...	...	...	...
13386	Más de 55	Universitario	Menos del 30%	Responsable Inscripto	3	4	0	3
13387	Más de 55	Universitario	Más del 50%	Responsable Inscripto	3	4	2	3
13388	Más de 55	Universitario	Entre 30% y 50%	Responsable Inscripto	3	4	1	3
13389	Más de 55	Universitario	Entre 30% y 50%	Responsable Inscripto	3	4	1	3
13390	Más de 55	Universitario	Entre 30% y 50%	Responsable Inscripto	3	4	1	3

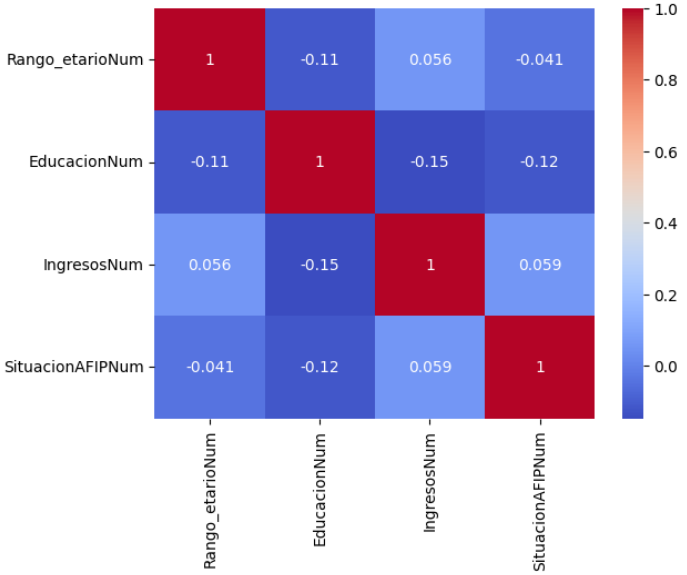
13391 rows × 8 columns

Aprovechamos para analizar la relación de algunas variables como edad y educación.

En este gráfico, se puede observar que en el rango etario ‘Más de 55 años’, hay una mayor proporción de individuos que sólo han completado la educación primaria, marcando una gran diferencia con el resto de los grupos etarios.



Generamos una matriz de correlación para ver cómo están vinculadas las variables elegidas.



Comenzamos a utilizar los métodos para clusterizar

K-Modes con Inicialización "CAO" y K-Modes con Inicialización "Huang"

K-Modes con Inicialización "Huang"

Aplicamos el método y nos tira la siguiente corrida

```
Initialization method and algorithm are deterministic. Setting n_init to 1.
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 0, cost: 25075.0
```

K-Modes con Inicialización "Huang"

```
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 0, cost: 25578.0
...
Run 2, iteration: 1/100, moves: 0, cost: 26049.0
...
Run 3, iteration: 1/100, moves: 2198, cost: 25120.0
...
Run 4, iteration: 1/100, moves: 1498, cost: 25010.0
...
Run 5, iteration: 1/100, moves: 0, cost: 26043.0
Best run was number 4
```

En cada corrida del algoritmo se registra el número de iteración actual, el número de movimientos realizados, el costo actual y el número de la ejecución. Se puede observar que en la ejecución número 4 se logró el menor costo, lo que indica que esa es la mejor solución encontrada por el algoritmo.

Resultados

Al comparar las corridas de K-Modes por los dos métodos de inicialización, encortramos los siguientes resultados:

```
Clusers generados por CAO
0    9639
1    3752
Name: cluster_CAO, dtype: int64

Clusers generados por Huang
0    9637
1    3754
Name: cluster_Huang, dtype: int64
```

Para llegar a que generen dos clusters de medidas similares, se utilizó un parámetro que regula la cantidad de corridas y se lo ajustó hasta obtener resultados similares en la cantidad de cada cluster tanto de CAO como de HUANG para poder compararlos.

Mostramos a continuación una tabla comparativa de las descripciones de los dos clusters generados de acuerdo al método utilizado.

Características de los clusters

Cluster mayor de CAO

```
count                  9639
unique                    4
top             Más de 55
freq                   4669
Name: Rango_etario, dtype: object

count                  9639
unique                    6
top             Secundario
freq                   4871
Name: Educación, dtype: object

count                  9639
unique                    3
top            Más del 50%
freq                   5363
Name: Ingresos, dtype: object

count                         9639
unique                          17
top     Responsable Inscripto
freq                          3742
Name: SituacionAFIP, dtype: object
```

Cluster mayor de HUANG

```
count                  9637
unique                    4
top             Más de 55
freq                   4617
Name: Rango_etario, dtype: object

count                  9637
unique                    6
top             Secundario
freq                   5073
Name: Educación, dtype: object

count                  9637
unique                    3
top            Más del 50%
freq                   5347
Name: Ingresos, dtype: object

count                         9637
unique                          17
top     Responsable Inscripto
freq                          3172
Name: SituacionAFIP, dtype: object
```

Cluster menor de CAO

```
count                  3752
unique                    4
top             De 31 a 45
freq                   2059
Name: Rango_etario, dtype: object
count                  3752
unique                    6
top              Primario
freq                   1993
Name: Educación, dtype: object
count                  3752
unique                    3
top             Entre 30% y 50%
freq                   2529
Name: Ingresos, dtype: object
count                        3752
unique                         17
top     Monotributo social
freq                          975
Name: SituacionAFIP, dtype: object
```

Cluster menor de HUANG

```
count                  3754
unique                    4
top             De 46 a 55
freq                   2031
Name: Rango_etario, dtype: object
count                  3754
unique                    6
top              Primario
freq                   2136
Name: Educación, dtype: object
count                  3754
unique                    3
top             Entre 30% y 50%
freq                   2623
Name: Ingresos, dtype: object
count                        3754
unique                         17
top     Responsable Inscripto
freq                          737
Name: SituacionAFIP, dtype: object
```

Podemos observar que en los clusters mayores obtenidos por cada método hay coincidencia en las características obtenidas, no así en el caso de los clusters con menor población, donde difieren en la edad y en la categoría de AFIP.

En cuanto a la edad, si bien difieren, son clases contiguas dentro de la variable, lo que no significa mayor distancia, y por lo tanto diferencia- entre ambos.

Con respecto a la diferencia en la categoría a la que pertenecen -o están inscriptos- en la AFIP, no encontramos un ordenamiento oficial que nos permita deducir mayores cosas con respecto a la "distancia" entre ambos resultados.

Utilizamos una técnica de reducción de dimensionalidad para visualizar los puntos en el espacio de dos dimensiones, con el Análisis de Componentes Principales (PCA) y el t-SNE.
Estas técnicas permiten proyectar los datos en un espacio de dos dimensiones de tal manera que se preserva la estructura de similitud entre los puntos en el espacio original.
Luego utilizamos un gráfico de dispersión para visualizar los puntos en el espacio de dos dimensiones y colorearlos según su etiqueta de cluster para ver si existe alguna estructura en los datos.

El primero corresponde a un gráfico de dos dimensiones utilizando PCA y el segundo con t-SNE.

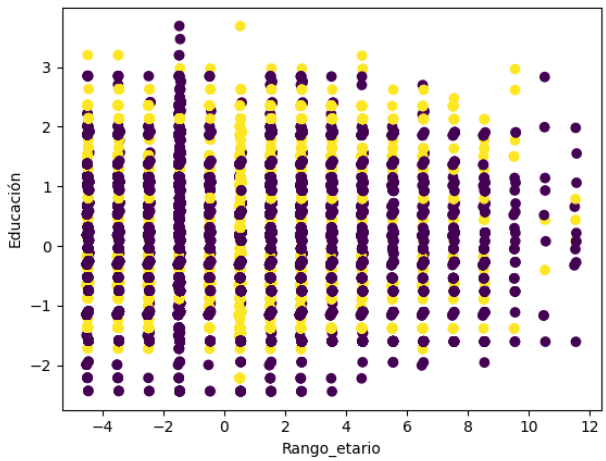


Gráfico PCA

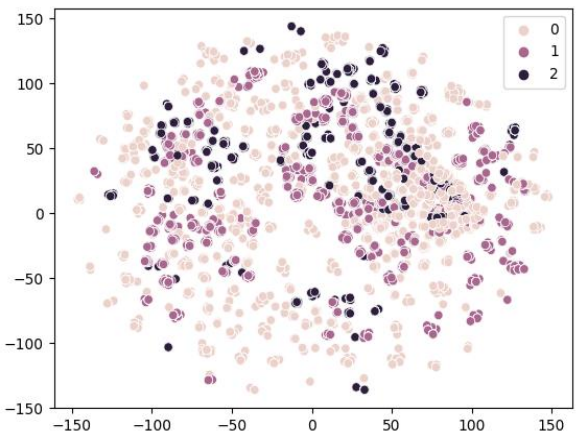


Gráfico t-SNE

En esta instancia, nos dimos cuenta que el ordenamiento de las categorías de AFIP quizás no era el correcto. Al modificarlo y volver a correr los datos, el K-Modes con CAO no sufrió ningún cambio, por el contrario, al usar Huang, los datos variaron bastante y descubrimos que; por cada corrida del algoritmo con los mismos parámetros, arrojaba, no sólo distintos cardinales de los clústers, sino distintas características de los mismos.

CONCLUSIONES

Con los resultados obtenidos llegamos a la conclusión de que hay que seguir investigando sobre el uso del K-Modes (CAO y Huang) y el comportamiento del mismo y de sus parámetros.

Si bien no pudimos sacar conclusiones firmes, en apariencia la utilización de K-Modes con CAO hace más estable el algoritmo, mostrando resultados que coinciden con el conocimiento empírico de la base de datos del Programa Cambio Rural, por lo que nos resulta satisfactorio el resultado general.

Material adicional:

Decisión Administrativa 1441/2020

DECAD-2020-1441-APN-JGM. Estructura organizativa

Incorporación CEyCR en la estructura del Ministerio

Funciones de la CEyCR

Anexo 4

Manual operativo

Notebook del proyecto

Referencias:

https://biblioguias.uma.es/citasybibliografia/ejemplosAPA

https://www.kaggle.com/code/halflingwizard/clustering-categorical-data-using-gower-distance

https://www.youtube.com/watch?v=S5cL5MAFon8

https://www.youtube.com/watch?v=o4bn2ZEGr4g

http://www.scielo.org.co/pdf/cide/v7n1/v7n1a03.pdf